G3 Genes | Genomes | Genetics

GENOME REPORT

Genome Sequence of the Banana Aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae) and Its Symbionts

Thomas C. Mathers,*,[1] Sam T. Mugford,* Saskia A. Hogenhout,* and Leena Tripathi[†,1]

*Department of Crop Genetics, John Innes Centre, Norwich Research Park, Norwich, United Kingdom and [†]International Institute of Tropical Agriculture (IITA), P.O. Box 30709-00100, Nairobi, Kenya

ORCID IDs: 0000-0002-8637-3515 (T.C.M.); 0000-0002-8537-5578 (S.T.M.); 0000-0003-1371-5606 (S.A.H.); 0000-0001-5723-4981 (L.T.)

ABSTRACT The banana aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae), is a major pest of cultivated bananas (*Musa* spp., order Zingiberales), primarily due to its role as a vector of *Banana bunchy top virus* (BBTV), the most severe viral disease of banana worldwide. Here, we generated a highly complete genome assembly of *P. nigronervosa* using a single PCR-free Illumina sequencing library. Using the same sequence data, we also generated complete genome assemblies of the *P. nigronervosa* symbiotic bacteria *Buchnera aphidicola* and *Wolbachia*. To improve our initial assembly of *P. nigronervosa* we developed a k-mer based deduplication pipeline to remove genomic scaffolds derived from the assembly of haplotigs (allelic variants assembled as separate scaffolds). To demonstrate the usefulness of this pipeline, we applied it to the recently generated assembly of the aphid *Myzus cerasi*, reducing the duplication of conserved BUSCO genes by 25%. Phylogenomic analysis of *P. nigronervosa*, our improved *M. cerasi* assembly, and seven previously published aphid genomes, spanning three aphid tribes and two subfamilies, reveals that *P. nigronervosa* falls within the tribe Macrosiphini, but is an outgroup to other Macrosiphini sequenced so far. As such, the genomic resources reported here will be useful for understanding both the evolution of Macrosphini and for the study of *P. nigronervosa*. Furthermore, our approach using low cost, high-quality, Illumina short-reads to generate complete genome assemblies of understudied aphid species will help to fill in genomic black spots in the diverse aphid tree of life.

KEYWORDS
Hemiptera
genome
assembly
insect vector
plant pest
phylogenomics

Aphids are economically important plant pests that cause damage to crops and ornamental plant species through parasitic feeding on plant sap and via the transmission of plant viruses. Of approximately 5,000 aphid species, around 100 have been identified as significant agricultural pests (Van Emden and Harrington 2017). Despite their economic importance, little to no genomic resources exist for many of these species or their relatives, hindering efforts to understand the evolution and ecology of aphid pests. To date, genome sequencing efforts have focused on members of the aphid tribe Macrosiphini (within subfamily Aphidinae), including the widely studied aphids *Acyrthosiphon pisum* (pea aphid) (International Aphid Genomics Consortium 2010; Li *et al.* 2019; Mathers *et al.* 2020) and *Myzus persicae* (green peach aphid) (Mathers *et al.* 2017, 2020), as well as other important pest species such as *Diuraphis noxia* (Russian wheat aphid) (Nicholson *et al.* 2015). Recently, additional genome sequences have become available for members of the tribe Aphidini (also in the subfamily Aphidinae) (Wenger *et al.* 2020; Thorpe *et al.* 2018; Chen *et al.* 2019; Quan *et al.* 2019; Mathers 2020) and the subfamily Lanchinae (Julca *et al.* 2020), broadening the phylogenetic scope of aphid genomic resources. However, many clades of the aphid phylogeny are still missing or underrepresented in genomic studies.

The banana aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae), is a major pest of cultivated bananas (*Musa* spp., order Zingiberales) and is widely distributed in tropical and subtropical

doi: https://doi.org/10.1534/g3.120.401358
Manuscript received May 6, 2020; accepted for publication October 1, 2020; published Early Online October 1, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.12251810.
[1]Corresponding authors: Department of Crop Genetics, John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK. E-mail: thomas.mathers@jic.ac.uk. International Institute of Tropical Agriculture (IITA), Nairobi, Kenya. E-mail: l.tripathi@cgiar.org

regions where bananas are grown (Waterhouse 1987). Like other aphid species, *P. nigronervosa* feeds predominantly from the phloem of its plant host. Intensive feeding can kill or affect the growth of young banana plants. However, direct feeding damage to adult plants is often negligible. Instead, the banana aphid causes most economic damage as a vector of plant viruses, some of which induce severe disease symptoms and substantial yield loss of banana (Dale 1987; Sharman *et al.* 2008; Savory and Ramakrishnan 2015). In particular, *P. nigronervosa* is the primary vector of the *Banana bunchy top virus* (BBTV), the most severe viral disease of banana worldwide (Dale 1987).

P. nigronervosa carries at least two bacterial symbionts: *Buchnera aphidicola* and *Wolbachia* (De Clerck *et al.* 2014). *Buchnera aphidicola* is an obligate (primary) symbiont present in almost all aphid species and provides essential amino acids to the aphids (Baumann 1995; Douglas 1998; Hansen and Moran 2011; Shigenobu and Wilson 2011). In contrast, *Wolbachia* is considered a facultative (secondary) symbiont and is found in a few aphid species at low abundance (Augustinos *et al.* 2011; Jones *et al.* 2011). Interestingly, *Wolbachia* is found systematically across the *P. nigronervosa* range (De Clerck *et al.* 2014) and is also present in the closely related species *P. caladii* van der Goot (Jones *et al.* 2011), which rarely colonizes banana, and prefers other plant species of the order Zingiberales (Foottit *et al.* 2010). Possibly, *Wolbachia* provides essential nutrients and vitamins to the *Pentalonia* spp or/and protects them from plant-produced defense molecules such as anti-oxidants or phenolic compounds of banana (Hosokawa *et al.* 2010).

Here, we generate highly complete genome assemblies of *P. nigronervos*a and its symbiotic bacteria *Buchnera aphidicola* and *Wolbachia,* using a single PCR-free Illumina sequencing library. Phylogenomic analysis reveals that *P. nigronervos*a falls within the aphid tribe Macrosiphini, but is an outgroup to other Macrosiphini sequenced so far. As such, the genomic resources reported here will useful for understanding the evolution of Macrosphini, and for the study of *P. nigronervos*a.

METHODS

Aphid rearing and sequencing library construction

A lab colony of *P. nigronervosa* was established from a single asexually reproducing female collected initially from the IITA's banana field at the International Livestock Research Institute (ILRI) Nairobi, Kenya. A single colony of *P. nigronervosa* was collected from a field-grown banana plant and introduced on an eight-week-old potted tissue culture banana plant in an insect-proof cage, placed in a glasshouse under room temperature and natural light. Pure aphid colonies were propagated by transferring a single aphid from the potted banana plant to another fresh young banana plant in the glasshouse every eight weeks. Aphids from this colony were used for all subsequent DNA and RNA extractions. Genomic DNA was extracted from a single individual with a modified CTAB protocol (based on Marzachi *et al.* 1998) and sent to Novogene (China), for library preparation and sequencing. Novogene prepared a PCR free Illumina sequencing library using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, USA), with the manufacturers protocol modified to give a 500 bp – 1 kb insert size. This library was sequenced on an Illumina HiSeq 2500 instrument with 250 bp paired-end chemistry. To aid scaffolding and genome annotation, we also generated a high coverage, strand-specific, RNA-seq library. RNA was extracted from whole bodies of 20-25 individuals using Trizol (Signma) followed by clean-up and on-column DNAse digestion using RNeasy (Qiagen) according to the manufactures' protocols, and sent to Novogene (China) where a sequencing library was prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, USA). This library was sequenced on an Illumina platform with 150 bp paired-end chemistry.

De novo genome assembly and quality control

Raw sequencing reads were processed with trim_galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore) to remove adapters and then assembled using Discovar *de novo* (https://software.broadinstitute.org/software/discovar/blog/) with default parameters. The content of this initial assembly was assessed with Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0 (Simão *et al.* 2015; Waterhouse *et al.* 2018) using the Arthropoda gene set (n = 1,066) and by k-mer analysis with the k-mer Analysis Toolkit (KAT) v2.2.0 (Mapleson *et al.* 2017), comparing k-mers present in the raw sequencing reads to k-mers found in the genome assembly with KAT comp. We identified a small amount of k-mer content that was present twice in the genome assembly but that had k-mer coverage in the reads of a single-copy region of the genome, indicating the assembly of haplotigs (allelic variants that are assembled into separate contigs) (Supplementary Figure 1a). To generate a close-to-haploid representation of the genome, we applied a strict filtering pipeline to the draft assembly based on k-mer analysis and whole genome self-alignment. First, the k-mer coverage of the homozygous portion of the genome was estimated with KAT distanalysis, which decomposes the k-mer spectra generated by KAT comp into discrete distributions corresponding to the number of times their content is found in the genome. Then, for each scaffold in the draft assembly, we used KAT sect to calculate the median k-mer coverage in the reads and the median k-mer coverage in the assembly. Scaffolds that had medium k-mer coverage of 2 in the assembly and median k-mer coverage in the reads that fell between the upper and lower bounds of homozygous genome content (identified by KAT distanalysis), were flagged as putative haplotigs. We then carried out whole genome self-alignment with nucmer v4.0.0beta2 (Marçais *et al.* 2018) and removed putative haplotigs that aligned to another longer scaffold in the genome with at least 75% identity and 25% coverage. The deduplicated assembly was then checked again with BUSCO and KAT comp to ensure that no (or minimal) genuine homozygous content had been lost from the assembly.

The deduplicated draft assembly was screened for contamination based on manual inspection of taxon-annotated GC content coverage plots ("blobplots") generated with BlobTools v1.0.1 (Kumar *et al.* 2013; Laetsch and Blaxter 2017). Genomic reads were aligned to the deduplicated draft assembly with BWA mem (Li 2013) and used to estimate average coverage per scaffold. Additionally, each scaffold in the assembly was compared to the NCBI nucleotide database (nt) with BLASTN v2.2.31 (Camacho *et al.* 2009). Read mappings and blast results were then passed to BlobTools which was used to create "blobplots" annotated with taxonomy at the order- and genus-level. Using this approach, we were able to identify and remove scaffolds corresponding to bacterial symbionts and scaffolds that had aberrant coverage and GC content patterns that are likely contaminants.

Finally, to further improve contiguity and gene-level completeness, we performed an additional round of scaffolding using our high coverage RNA-seq data with P_RNA_scaffodler (Zhu *et al.* 2018). RNA-seq reads were trimmed for adapters and low-quality bases with trim_galore and aligned to the deduplicated and cleaned assembly with HISAT2 v2.0.5 [-k 3 -pen-noncansplice 1000000] (Kim *et al.* 2015).

The resulting BAM file was then passed to P_RNA_scaffolder along with the draft assembly, and scaffolding performed with default settings. Gene-level completeness was assessed before and after RNA-seq scaffolding with BUSCO and final runs of KAT comp and BlobTools were performed to check the quality and completeness of the assembly.

Genome annotation

Repeats were identified and soft-masked in the frozen genome assembly using RepeatMasker v4.0.7 [-e ncbi -species insecta -a -xsmall -gff] (Smit *et al.* 2005) with the Repbase (Bao *et al.* 2015) Insecta repeat library. We then carried out gene prediction on the soft-masked genome using the BRAKER2 pipeline v2.0.4 (Lomsadze *et al.* 2014; Hoff *et al.* 2015) with RNA-seq evidence. BRAKER2 uses RNA-seq data to create intron hints and train a species-specific Augustus (Stanke *et al.* 2006, 2008) model which is subsequently used to predict protein coding genes, taking RNA-seq evidence into account. RNA-seq reads were aligned to the genome with HISAT2 v2.0.5 [–max-intronlen 25000–dta-cufflinks–rna-strandness RF] and the resulting BAM file passed to BRAKER2, which was run with default settings. Completeness of the BRAKER2 gene set was assessed using BUSCO with the Arthropoda gene set (n = 1,066). We generated a functional annotation of the predicted gene models using InterProScan v5.22.61 (Enright *et al.* 2002; Jones *et al.* 2014).

Upgrading Myzus cerasi v1.1

To demonstrate the usefulness of our k-mer based deduplication pipeline, we applied it to the published short-read assembly of *M. cerasi* (Mycer_v1.1) (Thorpe *et al.* 2018). We ran the pipeline as for *P. nigronervosa*, using the PCR-free Illumina reads that were originally used to assemble Mycer_v1.1 (NCBI bioproject PRJEB24287) and scaffolded the deduplicated assembly using RNA-seq data from Thorpe *et al.* (2016) (PRJEB9912) with P_RNA_scaffolder. RNA-seq reads were first trimmed for low quality bases and adapters with trim_galore, retaining reads where both members of a pair were at least 75 bp long after trimming. The deduplicated, scaffolded, assembly was ordered by size and assigned a numbered scaffold ID to create a frozen release for downstream analysis (Mycer_v1.2). Mycer_v1.2 was then soft-masked with RepeatMasker using the Repbase Insecta repeat library and protein coding genes predicted with BRAKER2 using the Thorpe *et al.* (2016) RNA-seq.

Phylogenomic analysis of aphids

Protein sequences from *P. nigronervosa*, our upgraded *M. cerasi* genome, and seven previously published aphid genomes (Supplementary Table 1), were clustered into orthogroups with OrthoFinder v2.2.3 (Emms and Kelly 2015, 2019). Where genes had multiple annotated transcripts, we used the longest transcript to represent the gene model. OrthoFinder is a comparative genomics pipeline that reconstructs orthogroups, estimates the rooted species tree, generates rooted gene trees, and infers orthologs and gene duplication events using the rooted gene trees, providing a rich resource for downstream comparative analysis. We ran OrthoFinder in multiple sequence alignment mode [-M msa -S diamond -T fasttree] using MAFFT (Katoh and Standley 2013) to align orthogroups and FastTree (Price *et al.* 2010) to infer maximum likelihood gene trees for each orthogroup. The species tree was then estimated based on a concatenated alignment of all conserved single-copy orthogroups and rooted using evidence from gene duplications with STRIDE (Emms and Kelly 2017). To confirm the topology recovered by the OrthoFinder–FastTree analysis, we carried out a bootstrapped maximum likelihood phylogenetic analysis based on the concatenated alignment with IQ-TREE v2.0.5 (Nguyen *et al.* 2015; Minh *et al.* 2020) and a coalescent analysis using conserved single copy gene trees with ASTRAL-III v5.6.3 (Mirarab *et al.* 2014; Mirarab and Warnow 2015; Zhang *et al.* 2018). For the IQ-TREE analysis, we automatically identified the optimum model of protein evolution with ModelFinder (Kalyaanamoorthy *et al.* 2017) and carried out 1,000 ultrafast bootstrap replicates (Hoang *et al.* 2018). For the ASTRAL-III analysis, we re-estimated gene trees for all conserved single-copy orthogroups using IQ-TREE with automatic protein model selection and ran ASTRAL-III with default settings.

Data availability

Sequence data and genome assemblies (including symbiont genomes) for this project have been deposited in NCBI databases under the project accession number PRJNA628023. The *P. nigronervosa* genome assembly and annotation, the updated *M. cerasi* genome assembly and annotation, orthogroup clustering results and code to run our assembly de-duplication pipeline are available for download from Zenodo (https://10.5281/zenodo.3765644). The *P. nigronervosa* genome assembly and annotation is also available from AphidBase (https://bipaa.genouest.org/sp/pentalonia_nigronervosa/). Supplemental material available at figshare: https://doi.org/10.25387/g3.12251810.

RESULTS AND DISCUSSION

P. nigronervosa genome assembly and annotation

In total we generated 23 Gb of PCR-free Illumina genome sequence data (~61x coverage of the *P. nigronervosa* genome) and 18 Gb of strand-specific RNA-seq data from a clonal lineage of *P. nigronervosa* (Supplementary Table 2). Using these data, we generated a *de novo* genome assembly of *P. nigronervosa* (Penig_v1). Penig_v1 is assembled into 18,348 scaffolds totaling 375 Mb of sequence with an N50 of 104 Kb (contig N50 = 64 Kb, n = 20,873; Table 1). The assembly is highly complete, with little duplicated or missing content (Figure 1a), and has excellent representation of conserved arthropod genes (95% complete and single-copy), meeting or exceeding the completeness of other published aphid genomes (Figure 1b). Furthermore, taxon annotated "blob-plots" show that Penig_v1 is free from obvious contamination (Supplementary Figure 2). Gene prediction using BRAKER2 with RNA-seq evidence resulted in the annotation of 27,698 protein coding genes and 29,708 transcripts. Completeness of the gene set reflects that of the genome assembly with 93% of BUSCO Arthropoda genes present as complete single copies in the annotation (Supplementary Figure 3). We were able to assign functional domains to 12,869 (47%) of the annotated gene models (Supplementary Table 3). Statistics for the final assembly and annotation of *P. nigronervosa* are summarized in Table 1.

P. nigronervosa in known to harbor the obligate aphid bacterial endosymbiont *Buchnera aphidicola* and a secondary symbiont, *Wolbachia*, that is found systematically across the species range (De Clerck *et al.* 2014). We identified both symbiotic bacteria in the initial discovar *de novo* assembly of *P. nigronervosa* (Figure 1c). *B. aphidicola* BPn was assembled into a single circular scaffold 617 KB in length, along with 2 plasmids. The *Wolbachia* WolPenNig assembly was more fragmented (1.46 Mb total length, 182 scaffolds, N50 = 15.5kb). Despite this, the WolPenNig assembly is likely highly complete as it is similar in size to both a more contiguous long-read assembly of a strain found in the soybean aphid (1.52 Mb total length,

Table 1 Genome assembly and annotation statistics for *P. nigronervosa* and *M. cerasi*

Species	*P. nigronervosa*	*M. cerasi*	*M. cerasi*
Assembly	Penig_v1	Mycer_v1.1	Mycer_v1.2
Base pairs (Mb)	375.35	405.71	393.23
% Ns	0.07	0.05	0.16
Number of contigs*	20,873	51,488	45,960
Contig N50 (Kb)*	64.06	19.7	20.6
Number of scaffolds	18,348	49,286	39,595
Scaffold N50 (Kb)	103.99	23.27	35.19
Longest scaffold (Kb)	631.82	265.36	350.78
Protein coding genes	27,698	28,688	31,070
Transcripts	29,708	28,688	33,159
Reference	This study	Thorpe *et al.* (2018)	This study

* Scaffolds split on runs of 10 or more Ns.

9 contigs, N50 = 841 Kb [Mathers 2020]) and to the reference assembly of *Wolbachia* wRi (Klasson *et al.* 2009) from *Drosophila simulans* (1.44 Mb total length, 1 contig). Furthermore, BUSCO analysis using the proteobacteria gene set (n = 221) reveals that WolPenNig has similar gene-level completeness to these high-quality assemblies, with 81% of BUSCO genes found as complete, single copies (Supplementary Figure 4).

Upgrading the Myzus cerasi genome assembly and annotation

The initial discovar *de novo* assembly of *P. nigronervosa* was moderately improved by applying our deduplication pipeline and by scaffolding the assembly with RNA-seq data. Compared to the raw discovar *de novo* assembly, contiguity increased by 8% (scaffold N50 = 104 kb *vs.* 96 Kb). Furthermore, the number of fragmented BUSCO

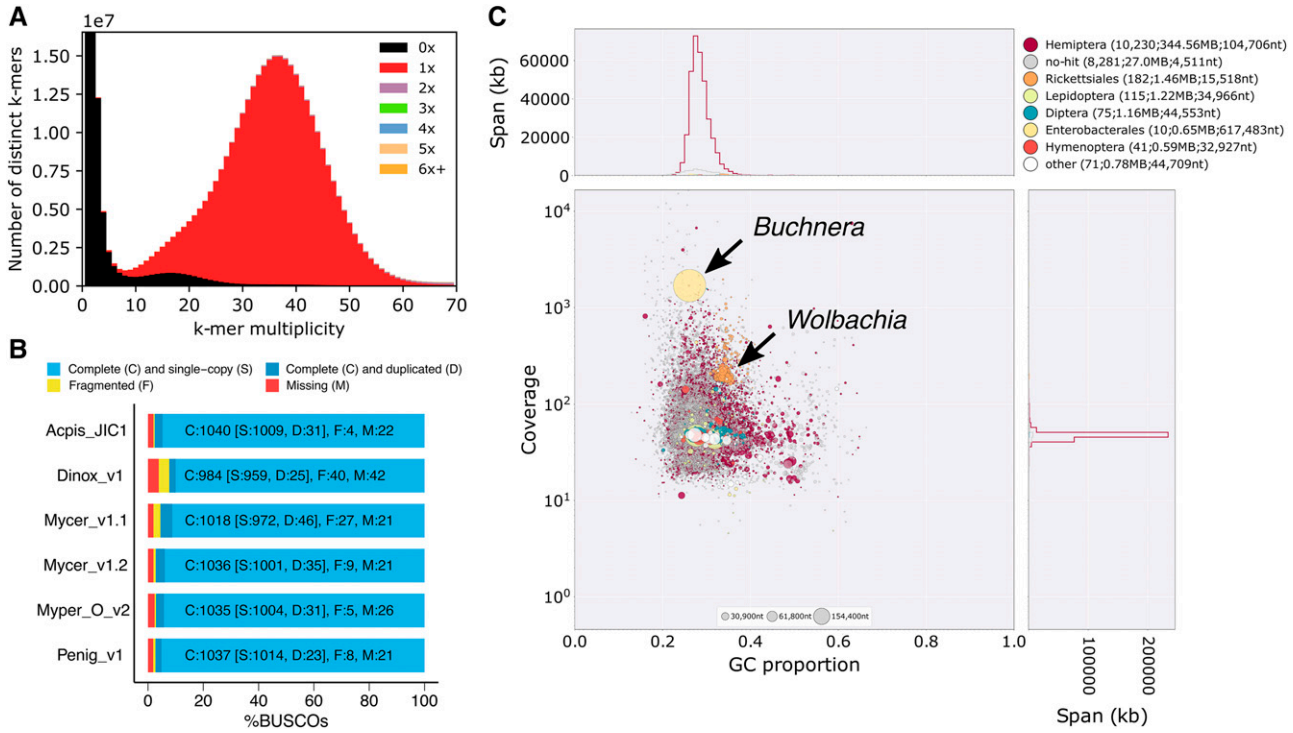

Figure 1 The *P. nigronervosa* genome assembly is complete and free from duplication and contamination. (a) KAT k-mer spectra plot comparing k-mer content of PCR-free *P. nigronervosa* Illumina reads to k-mer content of the final *P. nigronervosa* genome assembly (Penig_v1). Colors indicate how many times fixed length words (k-mers) from the reads appear in the assembly. Red indicates k-mers found only once in the assembly, black indicates content present in the reads but missing from the assembly and other colors indicate k-mers that are duplicated in the assembly. The x-axis shows the number of times each k-mer is found in the reads (k-mer multiplicity) and the y-axis shows the count of distinct k-mers in 1x k-mer multiplicity bins. (b) BUSCO analysis of Penig_v1, our updated assembly of *M. cerasi* (Mycer_v1.2) and published Macrosiphini genome assemblies. Myper_O_v2 = *Myzus persicae* clone O v2, Acpis_JIC1 = *Acyrthosiphon pisum* clone JIC1, Mycer_v1.1 = *Myzus cerasi* v1.1 and Dnox_v1 = *Diuraphis noxia* v1. The genomes were assessed using the Arthropoda gene set (n = 1,066). (c) Taxon-annotated GC content-coverage plot of the *P. nigronervosa* Discovar *de novo* genome assembly (post deduplication and prior to RNA-seq scaffolding – see Methods) showing co-assembly of the aphid and its symbionts. Each circle represents a scaffold in the assembly, scaled by length, and colored by order-level NCBI taxonomy assigned by BlobTools. The X axis corresponds to the average GC content of each scaffold and the Y axis corresponds to the average coverage based on alignment of *P. nigronervosa* PCR-free Illumina short reads. Marginal histograms show cumulative genome content (in Kb) for bins of coverage (Y axis) and GC content (X axis). Arrows highlight scaffolds assigned to the symbiotic bacteria *Buchnera aphidicola* and *Wolbachia* which were removed from the final assembly (Supplementary Figure 2).

Arthropoda genes was reduced from 11 to 8 indicating improved representation of the gene space in the processed assembly. Because the pipeline removes scaffolds that are predominantly made up of erroneously duplicated k-mers, these improvements were achieved without compromising genuine single-copy genome content (Supplementary Figure 1b). This approach will likely benefit other low-cost aphid genome assembly projects that use short-read sequencing, particularly when heterozygosity is high. To demonstrate this, we attempted to improve the published genome assembly of *Myzus cerasi* (Mcer_v1.1) (Thorpe *et al.* 2018), using publicly available data. Mcer_v1.1 is made up of 49,286 scaffolds, and k-mer analysis shows high heterozygosity and the presence duplicated content, likely the result of assembling haplotigs (Supplementary Figure 5a). We applied our deduplication and RNA-seq scaffolding pipeline to Mcer_v1.1 to create Mcer_v1.2. In total we removed 12.9 Mb of putatively duplicated content from Mcer_v1.1, reducing the assembly size from 405.5 to 392.6 Mb (Table 1). The updated assembly is 52% more contiguous than Mcer_v1.1 (scaffold N50 = 35 Kb *vs.* 23 Kb; Table 1) and BUSCO analysis indicates that Mcer_v1.2 better represents the gene space, with fewer duplicated (35 *vs.* 46) and fragmented (9 *vs.* 27) BUSCO Arthropoda genes (Figure 1b). As with Pnig_v1, these improvements were achieved without loss of genuine single-copy genome content (Supplementary Figure 5b). We annotated protein coding genes in Mcer_v1.2 with BRAKER2 using RNA-seq evidence, identifying 31,070 protein coding genes with 33,159 transcripts. Again, BUSCO analysis of the updated gene set indicates significant improvement over Mcer_v1.1, with the number of missing and fragmented BUSCO Arthropoda genes reduced from 65 to 20 and 55 to 20 respectively, and overall completeness increased by 8% from 946 to 1,026 BUSCO Arthropoda genes (Supplementary Figure 3).

P. nigronervosa is an outgroup to other sequenced Macrosiphini

To investigate the phylogenetic position of *P. nigronervosa* within aphids we carried out orthology clustering of 223,889 protein sequences from *P. nigronervosa*, our improved *M. cerasi* annotation, and seven previously published aphid genomes (Nicholson *et al.* 2015; Thorpe *et al.* 2018; Chen *et al.* 2019; Mathers 2020; Mathers *et al.* 2020). Although the number of aphid species with sequenced genomes is still low, the included species span three aphid tribes (Macrosphini, Aphidini and Lachnini) and approximately 100 million years of aphid evolution (Kim *et al.* 2011; Hardy *et al.* 2015; Julca *et al.* 2020). In total, 204,139 genes (85%) were clustered into 22,759 orthogroups, 4,721 of which are conserved and single-copy in all species (Supplementary table 4). Maximum likelihood phylogenetic analysis using a concatenated alignment of the single-copy orthogroups with FastTree produced a fully resolved species tree with 100% support at all nodes (Figure 2). The same fully supported topology was also recovered using maximum likelihood phylogenetic analysis with IQ-TREE (Supplementary Figure 6a) and when using the summary method ASTRAL-III (Supplementary Figure 6b), which performs well in the presence of incomplete lineage sorting (Mirarab *et al.* 2014). Macrosiphini and Aphidini are recovered as monophyletic groups in agreement with previous analyses based on a small number of genes (von Dohlen *et al.* 2006; Choi *et al.* 2018) and a recent phylogenomic analysis of aphids and other insects (Julca *et al.* 2020). *P. nigronervosa* is placed as an outgroup to other, previously sequenced, members of Macrosiphini (Figure 2).

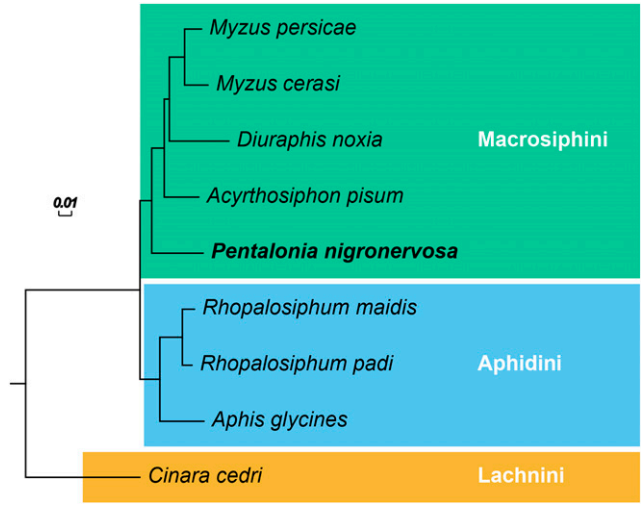

Figure 2 Maximum likelihood phylogeny of selected aphid species with sequenced genomes based on a concatenated alignment of 4,721 conserved one-to-one orthologs. All branches received maximal support based on the Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999) implemented in FastTree (Price *et al.* 2009, 2010) with 1,000 resamples. Clades are colored by aphid tribe. Branch lengths are in amino acid substitutions per site.

CONCLUSIONS

Using a single Illumina short-read sequence library and high-coverage RNA-seq data we have generated a high-quality draft genome assembly and annotation of the banana aphid and simultaneously assembled the genomes of its *Buchnera* and *Wolbachia* symbiotic bacteria, providing an important genomic resource for the future study of this important pest. Furthermore, as an outgroup to other sequenced aphids from the tribe Macrosiphini, the banana aphid genome will enable more detailed comparative analysis of a group that includes a large proportion of the most damaging aphid crop pests (Van Emden and Harrington 2017) as well as important model species such as the pea aphid (Brisson and Stern 2006) and the green peach aphid (Mathers *et al.* 2017, 2020).

ACKNOWLEDGMENTS

TCM is funded by a BBSRC Future Leader Fellowship (BB/R01227X/1). The described work was supported by a CEPAMs grant (17.03.2) to SH, a Bill and Melinda Gates Foundation grant (OPP1087428) awarded to LT, the BBSRC Institute Strategy Program (BB/P012574/1) award to the John Innes Centre, and the John Innes Foundation. This research was supported in part by the NBI Computing Infrastructure for Science Group, which provides technical support and maintenance to the John Innes Centre's high-performance computing cluster and storage systems.

LITERATURE CITED

Augustinos, A. A., D. Santos-Garcia, E. Dionyssopoulou, M. Moreira, A. Papapanagiotou *et al.*, 2011 Detection and characterization of Wolbachia infections in natural populations of Aphids: Is the hidden diversity fully unraveled? PLoS One 6: e28695. https://doi.org/10.1371/journal.pone.0028695

Bao, W., K. K. Kojima, and O. Kohany, 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6: 11. https://doi.org/10.1186/s13100-015-0041-9

Baumann, P., 1995 Genetics, physiology, and evolutionary relationships of the genus *Buchnera*: intracellular symbionts of aphids. Annu. Rev. Microbiol. 49: 55–94. https://doi.org/10.1146/annurev.mi.49.100195.000415

Brisson, J. A., and D. L. Stern, 2006 The pea aphid, *Acyrthosiphon pisum*: An emerging genomic model system for ecological, developmental and evolutionary studies. BioEssays 28: 747–755. https://doi.org/10.1002/bies.20436

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421. https://doi.org/10.1186/1471-2105-10-421

Chen, W., S. Shakir, M. Bigham, A. Richter, Z. Fei *et al.*, 2019 Genome sequence of the corn leaf aphid (*Rhopalosiphum maidis* Fitch). Gigascience 8: giz033. https://doi.org/10.1093/gigascience/giz033

Choi, H., S. Shin, S. Jung, D. J. Clarke, and S. Lee, 2018 Molecular phylogeny of Macrosiphini (Hemiptera: Aphididae): An evolutionary hypothesis for the Pterocomma-group habitat adaptation. Mol. Phylogenet. Evol. 121: 12–22. https://doi.org/10.1016/j.ympev.2017.12.021

De Clerck, C., T. Tsuchida, S. Massart, P. Lepoivre, F. Francis *et al.*, 2014 Combination of genomic and proteomic approaches to characterize the symbiotic population of the banana aphid (Hemiptera: Aphididae). Environ. Entomol. 43: 29–36. https://doi.org/10.1603/EN13107

Dale, J. L., 1987 Banana bunchy top: An economically important tropical plant virus disease. Adv. Virus Res. 33: 301–325. https://doi.org/10.1016/S0065-3527(08)60321-8

von Dohlen, C. D., C. A. Rowe, and O. E. Heie, 2006 A test of morphological hypotheses for tribal and subtribal relationships of Aphidinae (Insecta: Hemiptera: Aphididae) using DNA sequences. Mol. Phylogenet. Evol. 38: 316–329. https://doi.org/10.1016/j.ympev.2005.04.035

Douglas, A. E., 1998 Nutritional Interactions in Insect-Microbial Symbioses: Aphids and Their Symbiotic Bacteria Buchnera. Annu. Rev. Entomol. 43: 17–37. https://doi.org/10.1146/annurev.ento.43.1.17

Van Emden, H. F., and R. Harrington, 2017 Aphids as crop pests. Cab International, Wallingford, UK. https://doi.org/10.1079/9781780647098.0000

Emms, D. M., and S. Kelly, 2019 OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20: 238. https://doi.org/10.1186/s13059-019-1832-y

Emms, D. M., and S. Kelly, 2015 OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16: 157. https://doi.org/10.1186/s13059-015-0721-2

Emms, D. M., and S. Kelly, 2017 STRIDE: Species tree root inference from gene duplication events. Mol. Biol. Evol. 34: 3267–3278. https://doi.org/10.1093/molbev/msx259

Enright, A. J., S. Van Dongen, and C. A. Ouzounis, 2002 An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30: 1575–1584. https://doi.org/10.1093/nar/30.7.1575

Foottit, R. G., H. E. L. Maw, K. S. Pike, and R. H. Miller, 2010 The identity of Pentalonia nigronervosa Coquerel and P. caladii van der Goot (Hemiptera: Aphididae) based on molecular and morphometric analysis. Zootaxa 2358: 25–38. https://doi.org/10.11646/zootaxa.2358.1.2

Hansen, A. K., and N. A. Moran, 2011 Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. Proc. Natl. Acad. Sci. USA 108: 2849–2854. https://doi.org/10.1073/pnas.1013465108

Hardy, N. B., D. a. Peterson, and C. D. von Dohlen, 2015 The evolution of life cycle complexity in aphids: Ecological optimization or historical constraint? Evolution (N. Y.) 69: 1423–1432.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh, 2018 UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35: 518–522. https://doi.org/10.1093/molbev/msx281

Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, 2015 BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767–769. https://doi.org/10.1093/bioinformatics/btv661

Hosokawa, T., R. Koga, Y. Kikuchi, X. Y. Meng, and T. Fukatsu, 2010 Wolbachia as a bacteriocyte-associated nutritional mutualist. Proc. Natl. Acad. Sci. USA 107: 769–774. https://doi.org/10.1073/pnas.0911476107

International Aphid Genomics Consortium, 2010 Genome sequence of the pea aphid *Acyrthosiphon pisum.* PLoS Biol. 8: e1000313. https://doi.org/10.1371/journal.pbio.1000313

Jones, P., D. Binns, H. Y. Chang, M. Fraser, W. Li *et al.*, 2014 InterProScan 5: Genome-scale protein function classification. Bioinformatics 30: 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Jones, R. T., A. Bressan, A. M. Greenwell, and N. Fierer, 2011 Bacterial communities of two parthenogenetic aphid species cocolonizing two host plants across the Hawaiian islands. Appl. Environ. Microbiol. 77: 8345–8349. https://doi.org/10.1128/AEM.05974-11

Julca, I., M. Marcet-houben, F. Cruz, C. Vargas-chavez, J. Spencer *et al.*, 2020 Phylogenomics identifies an ancestral burst of gene duplications predating the diversification of Aphidomorpha. Mol. Biol. Evol. 37: 730–756. https://doi.org/10.1093/molbev/msz261

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. Von Haeseler, and L. S. Jermiin, 2017 ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14: 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30: 772–780. https://doi.org/10.1093/molbev/mst010

Kim, D., B. Langmead, and S. L. Salzberg, 2015 HISAT: A fast spliced aligner with low memory requirements. Nat. Methods 12: 357–360. https://doi.org/10.1038/nmeth.3317

Kim, H., S. Lee, and Y. Jang, 2011 Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): diversification, host association, and biogeographic origins. PLoS One 6: e24749. https://doi.org/10.1371/journal.pone.0024749

Klasson, L., J. Westberg, P. Sapountzis, K. Näslund, Y. Lutnaes *et al.*, 2009 The mosaic genome structure of the Wolbachia wRi strain infecting Drosophila simulans. Proc. Natl. Acad. Sci. USA 106: 5725–5730. https://doi.org/10.1073/pnas.0810753106

Kumar, S., M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter, 2013 Blobology: exploring raw genome data for contaminants, symbionts, and parasites using taxon-annotated GC-coverage plots. Front. Genet. 4: 1–12. https://doi.org/10.3389/fgene.2013.00237

Laetsch, D. R., and M. L. Blaxter, 2017 BlobTools: Interrogation of genome assemblies. F1000 Res. 6: 1287. https://doi.org/10.12688/f1000research.12232.1

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997. https://arxiv.org/abs/1303.3997v2

Li, Y., H. Park, T. E. Smith, and N. A. Moran, 2019 Gene family evolution in the pea aphid based on chromosome-level genome assembly. Mol. Biol. Evol. 36: 2143–2156. https://doi.org/10.1093/molbev/msz138

Lomsadze, A., P. D. Burns, and M. Borodovsky, 2014 Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 42: e119. https://doi.org/10.1093/nar/gku557

Mapleson, D., G. G. Accinelli, G. Kettleborough, J. Wright, and B. J. Clavijo, 2017 KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics 33: 574–576.

Marçais, G., A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg *et al.*, 2018 MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. 14: e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Marzachi, C., F. Veratti, and D. Bosco, 1998 Direct PCR detection of phytoplasmas in experimentally infected insects. Ann. Appl. Biol. 133: 45–54. https://doi.org/10.1111/j.1744-7348.1998.tb05801.x

Mathers, T. C., 2020 Improved genome assembly and annotation of the soybean aphid (*Aphis glycines* Matsumura). G3 (Bethesda) 10: 899–906.

Mathers, T. C., Y. Chen, G. Kaithakottil, F. Legeai, S. T. Mugford *et al.*, 2017 Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. Genome Biol. 18: 27. https://doi.org/10.1186/s13059-016-1145-3

Mathers, T. C., R. H. M. Wouters, S. T. Mugford, D. Swarbreck, C. van Oosterhout *et al.*, 2020 Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome. Mol. Biol. Evol.. https://doi.org/10.1093/molbev/msaa246

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams *et al.*, 2020 IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37: 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson *et al.*, 2014 ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30: i541–i548. https://doi.org/10.1093/bioinformatics/btu462

Mirarab, S., and T. Warnow, 2015 ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31: i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Nguyen, L. T., H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, 2015 IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32: 268–274. https://doi.org/10.1093/molbev/msu300

Nicholson, S. J., M. L. Nickerson, M. Dean, Y. Song, P. R. Hoyt *et al.*, 2015 The genome of *Diuraphis noxia*, a global aphid pest of small grains. BMC Genomics 16: 429. https://doi.org/10.1186/s12864-015-1525-1

Price, M. N., P. S. Dehal, and A. P. Arkin, 2009 FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26: 1641–1650. https://doi.org/10.1093/molbev/msp077

Price, M. N., P. S. Dehal, and A. P. Arkin, 2010 FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490. https://doi.org/10.1371/journal.pone.0009490

Quan, Q., X. Hu, B. Pan, B. Zeng, N. Wu *et al.*, 2019 Draft genome of the cotton aphid *Aphis gossypii*. Insect Biochem. Mol. Biol. 105: 25–32. https://doi.org/10.1016/j.ibmb.2018.12.007

Savory, F. R., and U. Ramakrishnan, 2015 Cryptic diversity and habitat partitioning in an economically important aphid species complex. Infect. Genet. Evol. 30: 230–237. https://doi.org/10.1016/j.meegid.2014.12.020

Sharman, M., J. E. Thomas, S. Skabo, and T. A. Holton, 2008 Abacá bunchy top virus, a new member of the genus Babuvirus (family Nanoviridae). Arch. Virol. 153: 135–147. https://doi.org/10.1007/s00705-007-1077-z

Shigenobu, S., and A. C. C. Wilson, 2011 Genomic revelations of a mutualism: The pea aphid and its obligate bacterial symbiont. Cell. Mol. Life Sci. 68: 1297–1309. https://doi.org/10.1007/s00018-011-0645-2

Shimodaira, H., and M. Hasegawa, 1999 Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol. Biol. Evol. 16: 1114–1116. https://doi.org/10.1093/oxfordjournals.molbev.a026201

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, A. F. A., R. Hubley, and P. Green, 2005 RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org

Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler, 2008 Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24: 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stanke, M., O. Schöffmann, B. Morgenstern, and S. Waack, 2006 Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7: 62. https://doi.org/10.1186/1471-2105-7-62

Thorpe, P., P. J. A. Cock, and J. Bos, 2016 Comparative transcriptomics and proteomics of three different aphid species identifies core and diverse effector sets. BMC Genomics 17: 172. https://doi.org/10.1186/s12864-016-2496-6

Thorpe, P., C. M. Escudero-Martinez, P. J. A. A. Cock, S. Eves-Van Den Akker, J. I. B. B. Bos *et al.*, 2018 Shared transcriptional control and disparate gain and loss of aphid parasitism genes. Genome Biol. Evol. 10: 2716–2733. https://doi.org/10.1093/gbe/evy183

Waterhouse, D. F., 1987 Pentalonia nigronervosa Coquerel, pp. 42–49 in *Biological Control: Pacific Prospects*, edited by Waterhouse, D. F., and K. R. Norris. Inkata Press, Melbourne.

Waterhouse, R. M., M. Seppey, F. A. Simao, M. Manni, P. Ioannidis *et al.*, 2018 BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35: 543–548. https://doi.org/10.1093/molbev/msx319

Wenger, J. A., B. J. Cassone, F. Legeai, J. S. Johnston, R. Bansal *et al.*, 2020 Whole genome sequence of the soybean aphid, *Aphis glycines*. Insect Biochem. Mol. Biol. 123: 102917. https://doi.org/10.1016/j.ibmb.2017.01.005

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab, 2018 ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19: 153. https://doi.org/10.1186/s12859-018-2129-y

Zhu, B. H., J. Xiao, W. Xue, G. C. Xu, M. Y. Sun *et al.*, 2018 P_RNA_scaffolder: A fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics 19: 175. https://doi.org/10.1186/s12864-018-4567-3

Communicating editor: R. Kulathinal